Lecture
"Foundations of Language Science and Technology: Semi-Supervised Learning"

Prof. Dr. D. Klakow

Exercise

Please answer the following theoretical questions on semi-supervised learning. All questions should be answered within no more than two or three sentences.

Send your answers to `Michael.Wiegand@lsv.uni-saarland.de`

The **deadline** for this exercise is 12/12/2008 8:30am.

The answers will be discussed in the exercise session on 12/12/2008.

1. Why is semi-supervised learning useful?
2. Name a semi-supervised learning algorithm that has been explained in the lecture and explain why it can be regarded as a bootstrapping algorithm.
3. Explain why a semi-supervised classifier heavily relies on the correlation between features observed in the labeled data instances of a specific class c_i and the other features only observed in the unlabeled data instances of c_i.
4. Explain how a semi-supervised classifier might be led astray.
5. Why is feature selection more important in semi-supervised learning for text classification than in supervised learning (where a sufficient amount of labeled documents is available)?
6. Give a list of possible parameters that have to be taken into account in semi-supervised learning.
7. With regard to the amount of labeled training data, there are two (extreme) situations in which semi-supervised learning does not work. Name them.
8. Imagine, you are to implement an EM-classifier for spam classification. Your entire dataset comprises 1000 spam and ham mails each. You are to use 1 labeled document per class only. In your first version you only define all words you observe in the labeled documents as your overall vocabulary. You get very bad results. Explain why your current choice of the vocabulary is inappropriate? What would you suggest as an alternative vocabulary?

9. One of your eager fellow students has done some experiments in semi-supervised learning on a standard dataset for binary text classification. As a feature set he manually compiled a list of words which he thinks are very discriminative for this classification task. He spent *three weeks* building this resource. With the new lexicon, he achieves a better classification accuracy than a supervised classifier trained on the labeled data (50 documents per class) only. He tells you that he is now convinced that semi-supervised learning is superior to supervised learning. You do not share his enthusiasm. Explain why!